Using 3D affordance for grasping pose generation

BY JOHANNES KOLHOFF

Outline

Motivation

What is Affordance?

Background

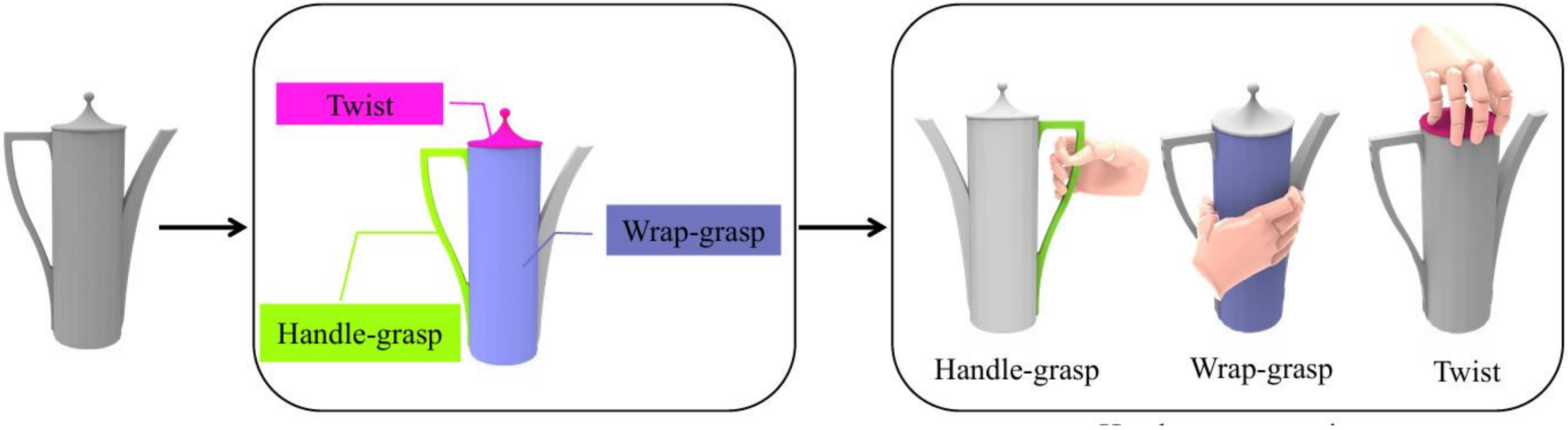
Current models:

- Implicit Estimation and Visual Affordance
- Language-Conditioned Affordance-Pose Detection
- Language Guided Affordance (AffordDexGrasp)

Summary

Motivation

Where would you grasp this teapot?



[2]

What is Affordance

„Affordance“ depends on:

Object shape

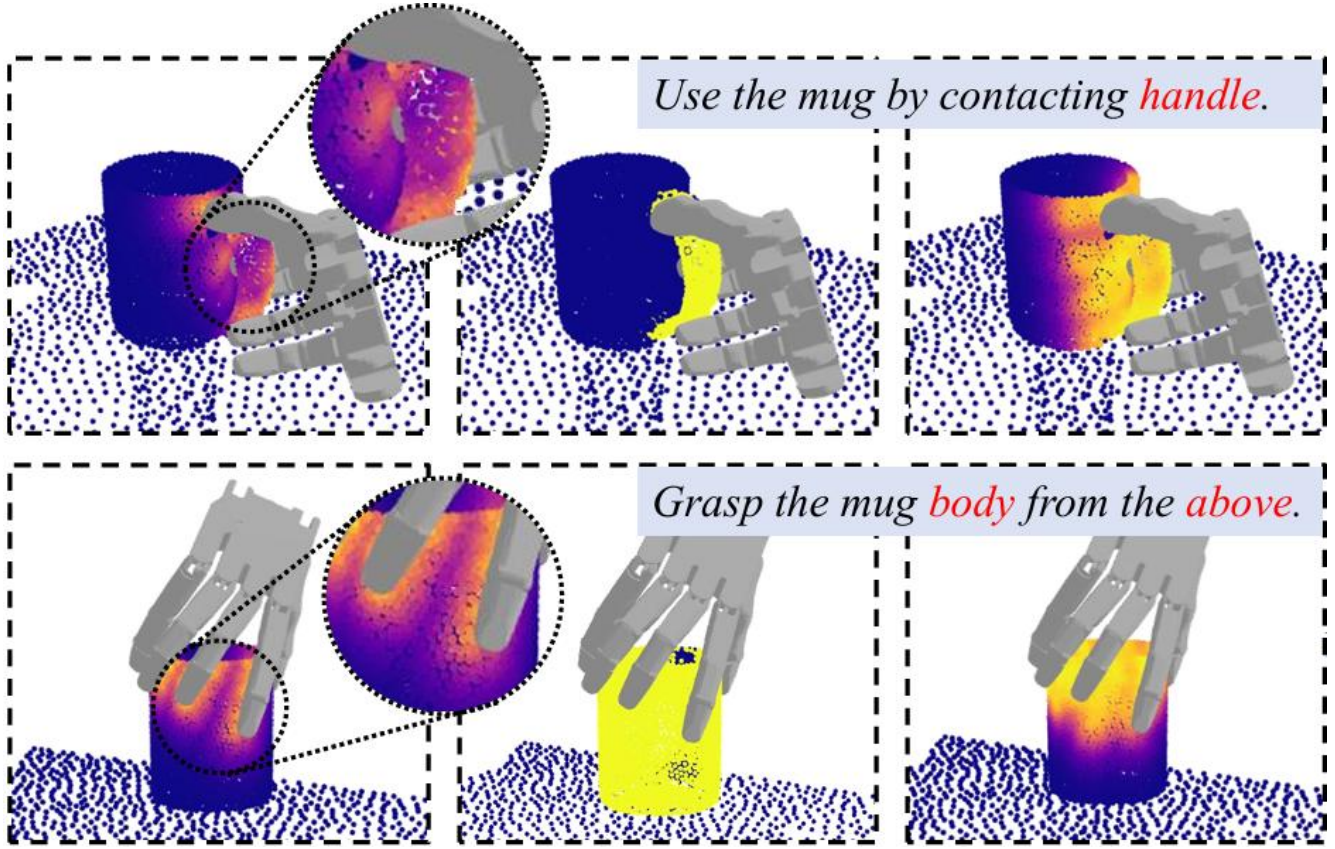
Object type (what is it)

Intention (what do you want to do?)

Also, your manipulator

=> Object in its context!

How an object can be used



[5]

Challenges:

- Understanding the objects
- Unknown objects
- Many affordance category
- Bridge to 6DoF grasp pose
- Bridge from language guided
- Cluttered environments (pile)
- Datasets

Generalizability

Old and new approaches

LEGACY APPROACHES	STILL IN USE	NEW RESEARCH
2D images	Depth cameras	6-DoF grasp pose estimation
Bounding boxes	Point clouds	Language guided
Masks	Pixel/point-based ground truths	Neural networks
Semantic recognition/ segmentation	Random grasp generation	Some combination of affordance and pose generation(“End-to-End“)

Implicit Estimation and Visual Affordance

Input: point cloud (depth camera render)

Parallel processing of affordance and grasp detection

- Combination at the end

Defined “mini-task actions” (set of affordances)

- E.g. wrap, grasp, pour, cut etc.

Confidence scores and affordance map are used to determine gasp candidate

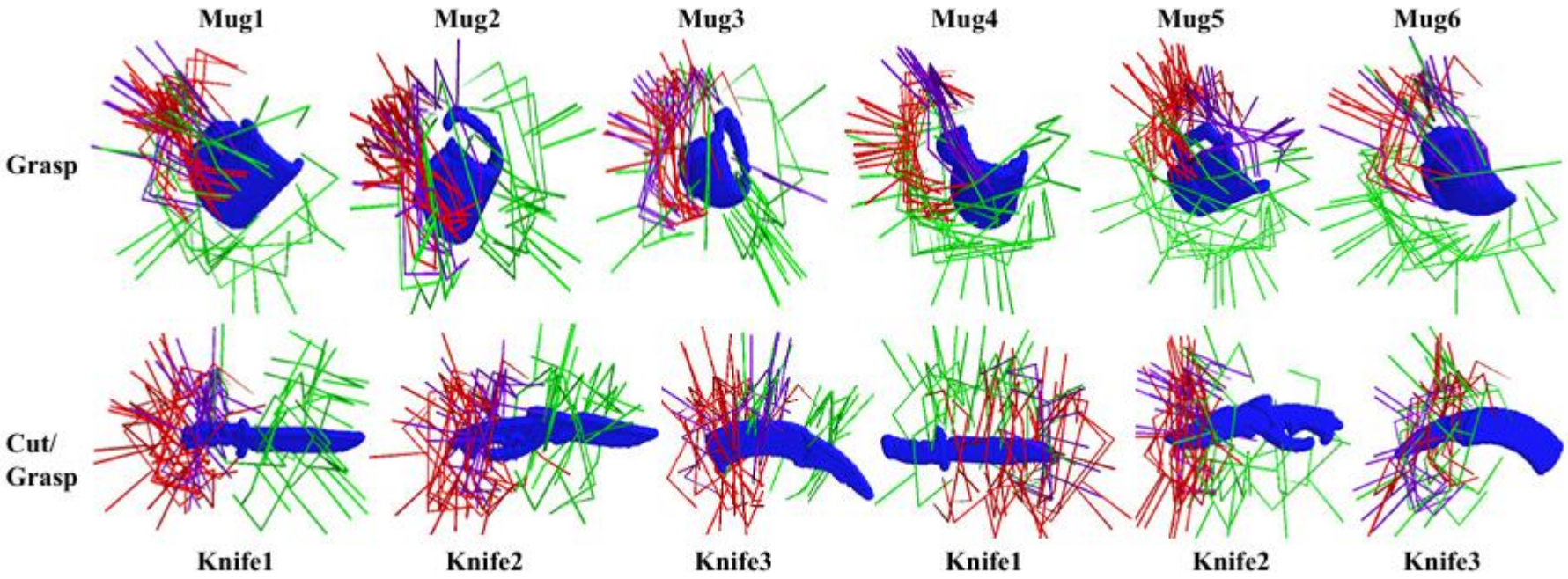
[3]

Implicit Estimation and Visual Affordance

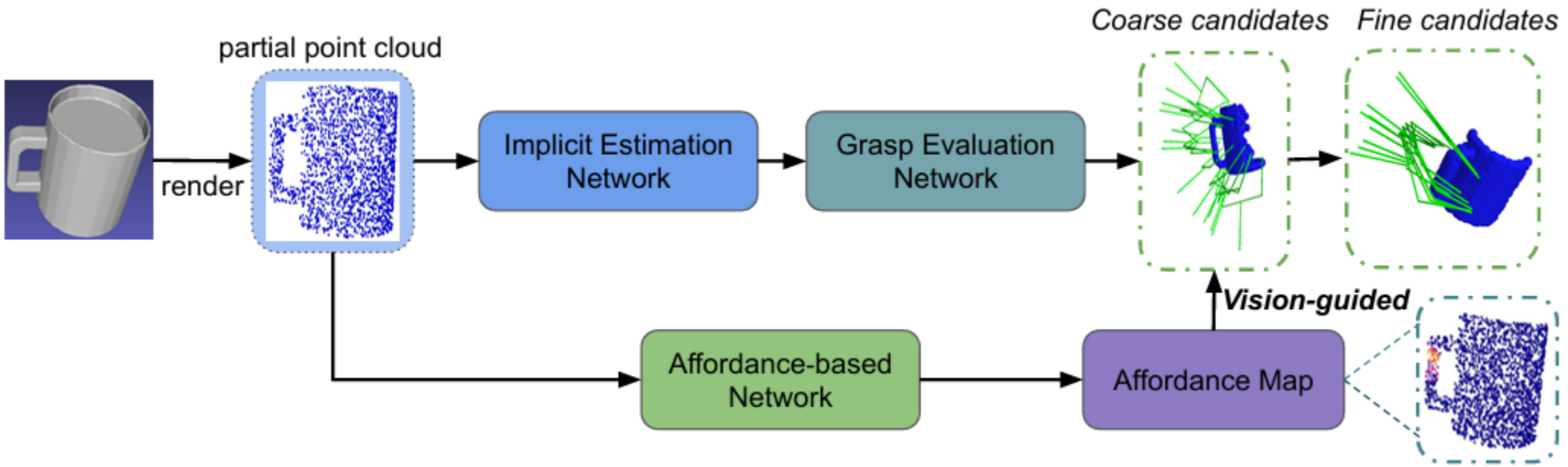
new dataset: simulated ShapeNet with task-oriented category's

Labeling: binary (successful or not)

Red: ground truth



Implicit Estimation and Visual Affordance



[3]

Implicit Estimation and Visual Affordance

ADVANTAGE

Only image (point cloud) as input

Can evaluate and determine different affordances

DISADVANTAGE

Closed set

No natural language guidance

Not easily expandable to new objects/affordances

Language-Conditioned Affordance-Pose Detection - Dataset

New dataset based on 3D AffordanceNet and 6-DoF GraspNet

Manual annotation of affordance category’s

Converted to point cloud

Dataset is triplet of point could, affordance category and pose

[4]

Language-Conditioned Affordance-Pose Detection

Input: point cloud and text (affordance)

Point cloud and text are encoded separately

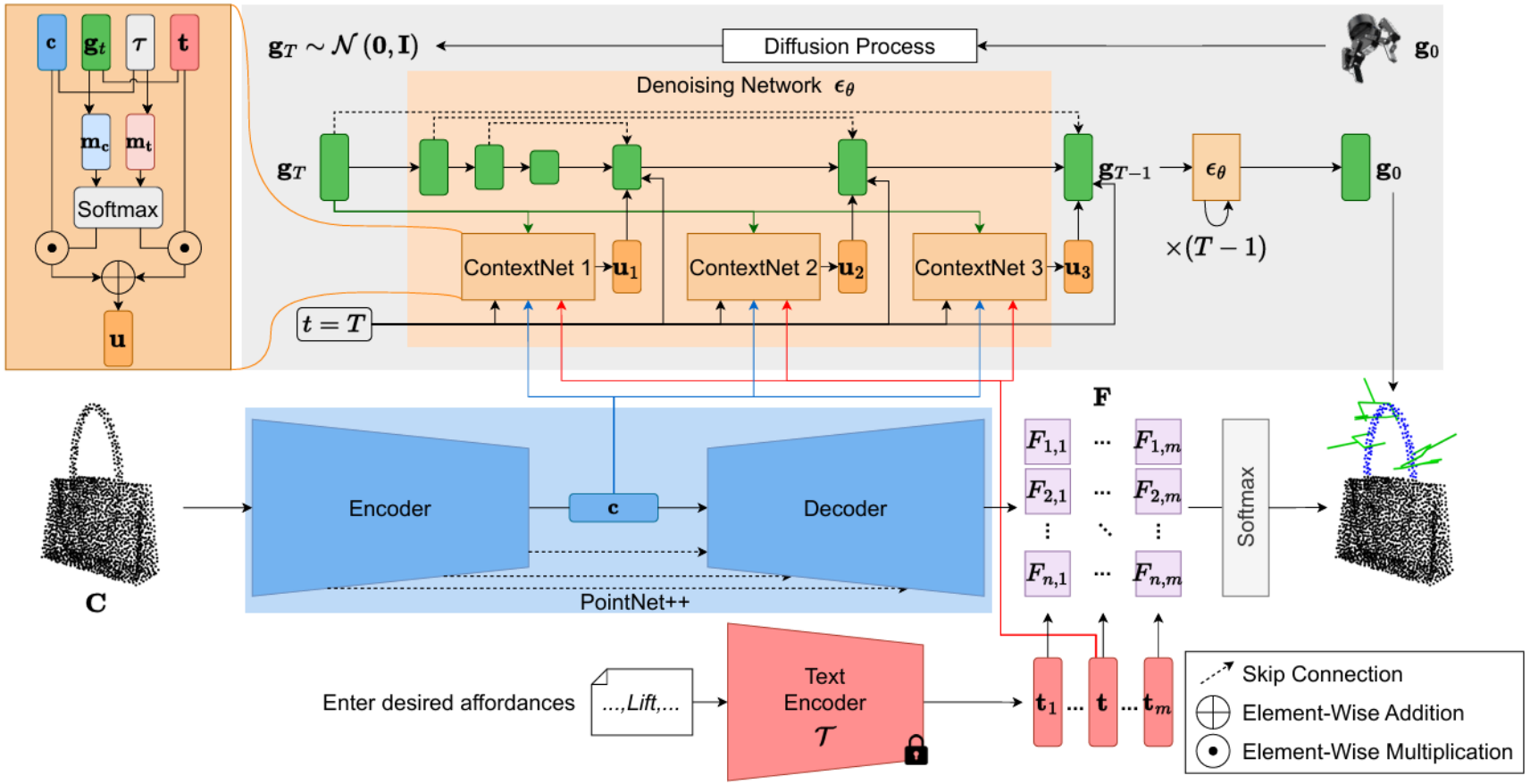
Grasp estimation is diffusion based

Context for diffusion is made up of embeddings of:

- Point cloud, text, robot state and timestep

[4]

Language-Conditioned Affordance-Pose Detection



[4]

Language-Conditioned Affordance-Pose Detection

ADVANTAGE

Open vocabulary

Expansion to new affordances and objects possible

DISADVANTAGE

Always requires affordance instruction

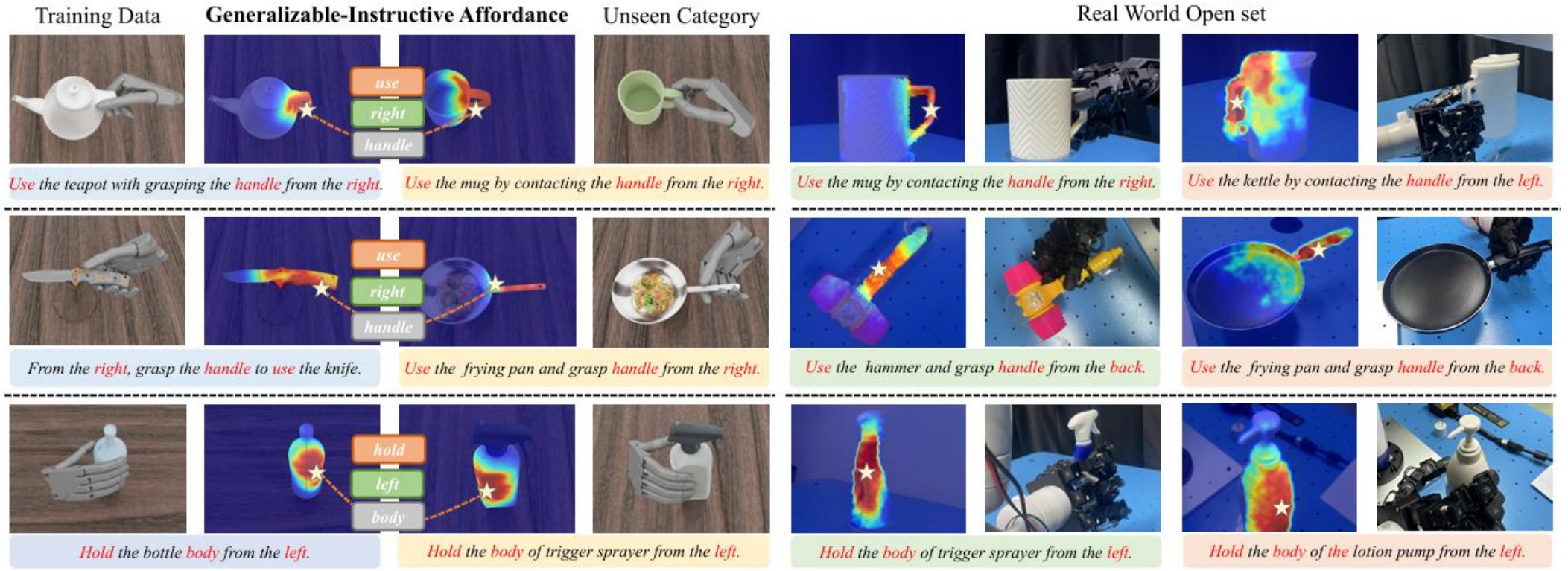
Expansion to new affordances limited

Word choice matters

Single object only

AffordDexGrasp: Dataset

New open set dataset based on language-guided dexterous grasp dataset
tabletop environment and 33 categories



[5]

AffordDexGrasp:

Input:

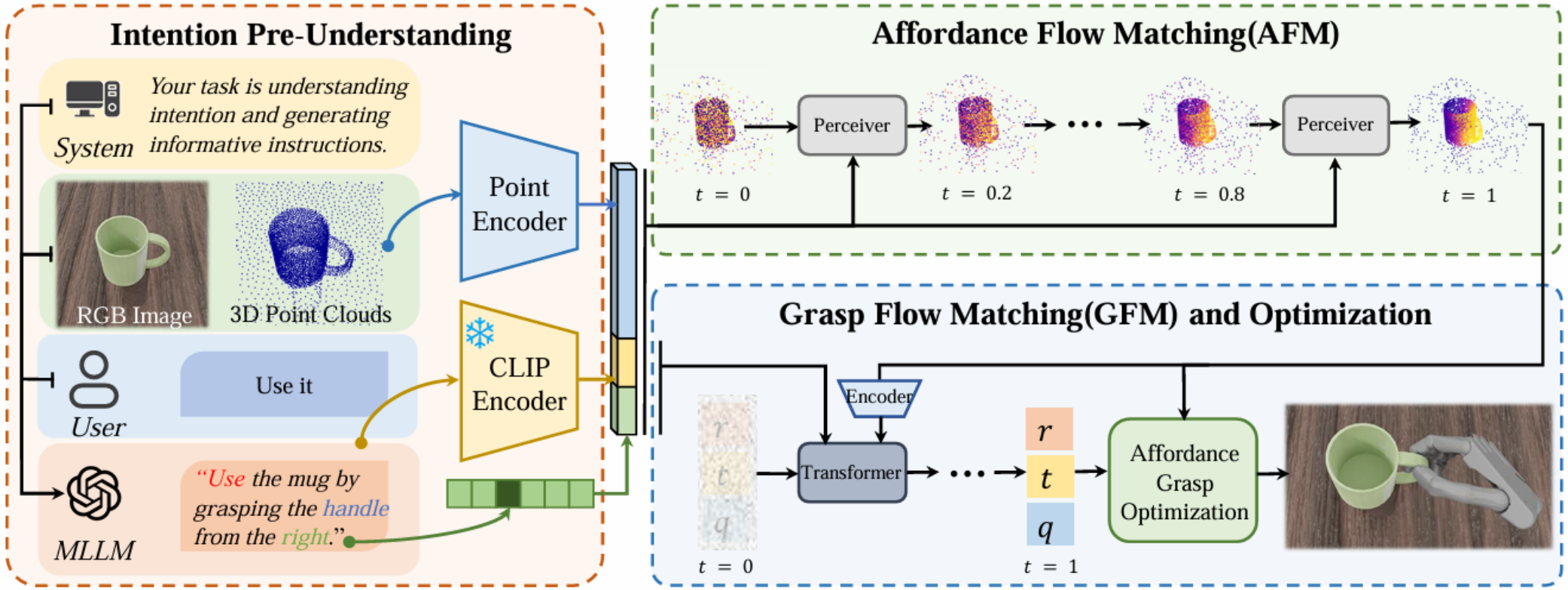
- Natural language user input
- RGB image
- Point cloud

Utilizes MLLM to intuit and guide affordance

Closer to sequential processing:

- First create an affordance map
- Then determine gasp from affordance map + affordance encoding

AffordDexGrasp:



AffordDexGrasp:

ADVANTAGE

Open to natural language instructions
- But not necessary!

Use of RGB image

Robust to different formulations

DISADVANTAGE

Input needs RBG image, point cloud and instruction for best performance

Still has real world gap

Summary

Utilize 3D/2.5D vision

Gather context information (image + text/situation)

Integration of affordance and grasp pose estimation

Challenges:

- Pressley defining the problem (+dataset)
- Measuring success
- Applying to real world (multiple objects, point cloud render etc.)

Sources

[1] “Visual Affordance and Function Understanding: A Survey“ 2018 Mohammed Hassanin, Salman Khan, Murat Tahtali

[2] “AffordPose: A Large-scale Dataset of Hand-Object Interactions with Affordance-driven Hand Pose” 2023 Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, Jian Liu

[3] “Learning 6-DoF Task-oriented Grasp Detection via Implicit Estimation and Visual Affordance” 2022 Wenkai Chen, Hongzhuo Liang, Zhaopeng Chen, Fuchun Sun and Jianwei Zhang

[4] “Language-Conditioned Affordance-Pose Detection in 3D Point Clouds” 2024 Toan Nguyen, Minh Nhat Vu, Baoru Huang, Tuan Van Vo, Vy Truong, Ngan Le Thieu Vo, Bac Le, Anh Nguyen

[5] “AffordDexGrasp: Open-set Language-guided Dexterous Grasp with Generalizable-Instructive Affordance” 2025 Yi-Lin Wei, Mu Lin, Yuhao Lin, Jian-Jian Jiang, Xiao-Ming Wu, Ling-An Zeng, Wei-Shi Zheng